Genomic Evolution of the Ascomycete Yeasts

Robert Riley[1], Sajeet Haridas[1], Asaf Salamov[1], Kyria Boundy-Mills[2], Markus Goker[3], Chris Hittinger[4], Hans-Peter Klenk[5], Mariana Lopes[4], Jan P. Meir-Kolthoff[3], Antonis Rokas[6], Carlos Rosa[7], Carmen Scheuner[3], Marco Soares[4], Benjamin Stielow[8], Jennifer H. Wisecaver[6], Ken Wolfe[9], Meredith Blackwell[10], Cletus Kurtzman[11], Igor Grigoriev[1], Thomas Jeffries[12]

[1]US Department of Energy Joint Genome Institute, Walnut Creek, CA
[2]Department of Food Sciences and Technology, University of California Davis, Davis, CA
[3]Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany
[4]Laboratory of Genetics, Genetics/ Biotechnology Center, Madison, WI
[5]School of Biology, Newcastle University, Newcastl upon Tyne, UK
[6]Department of Biological Sciences, Vanderbilt University
[7]Instituto de Ciencias Biologicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[8]CBS-KNAW Fungal Biodiversity Centre, Utrecht, Netherlands
[9]UCD School of Medicine & Medical Science, Conway Institute, University College Dublin, Dublin, Ireland
[10]Department of Biological Sciences, Lousiana State University, Baton Rouge, LA
[11]USDA ARS, MWA, NCAUR, BFPM, Peoria, Il
[12]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI

March 2015

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

LBNL-

Genomic evolution of the ascomycete yeasts

Robert Riley[1], Sajeet Haridas[1], Asaf Salamov[1], Kyria Boundy-Mills[2], Markus Göker[3], Chris Hittinger[4], Hans-Peter Klenk[5], Mariana Lopes[4], Jan P. Meier- Kolthoff[3], Antonis Rokas[6], Carlos Rosa[7], Carmen Scheuner[3], Marco Soares[4], Benjamin Stielow[8], Jennifer H. Wisecaver[6], Ken Wolfe[9], Meredith Blackwell[10], Cletus Kurtzman[11], Igor Grigoriev[1], Thomas Jeffries[12]

1) US Department of Energy Joint Genome Institute, Walnut Creek, CA; 2) Department of Food Science and Technology, University of California Davis, Davis, CA; 3) Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Braunschweig, GERMANY; 4) Laboratory of Genetics, Genetics/Biotechnology Center, Madison, WI; 5) School of Biology, Newcastle University, Newcastle upon Tyne, UK; 6) Department of Biological Sciences, Vanderbilt University; 7) Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil; 8) CBS-KNAW Fungal Biodiversity Centre, Utrecht, Netherlands; 9) UCD School of Medicine & Medical Science, Conway Institute, University College Dublin, Dublin, Ireland; 10) Department of Biological Sciences, Louisiana State University, Baton Rouge, LA; 11) USDA, ARS, MWA, NCAUR, BFPM, Peoria, IL; 12) Department of Bacteriology, University of Wisconsin-Madison, Madison, WI

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

Abstract

Yeasts are important for industrial and biotechnological processes and show remarkable metabolic and phylogenetic diversity despite morphological similarities. We have sequenced the genomes of 16 ascomycete yeasts of taxonomic and industrial importance including members of Saccharomycotina and Taphrinomycotina. Phylogenetic analysis of these and previously published yeast genomes helped resolve the placement of species including *Saitoella complicata*, *Babjeviella inositovora*, *Hyphopichia burtonii*, and *Metschnikowia bicuspidata*. Moreover, we find that alternative nuclear codon usage, where CUG encodes serine instead of leucine, are monophyletic within the Saccharomycotina. Most of the yeasts have compact genomes with a large fraction of single exon genes, and a tendency towards more introns in early-diverging species. Analysis of enzyme phylogeny gives insights into the evolution of metabolic capabilities such as methanol utilization and assimilation of alternative carbon sources.

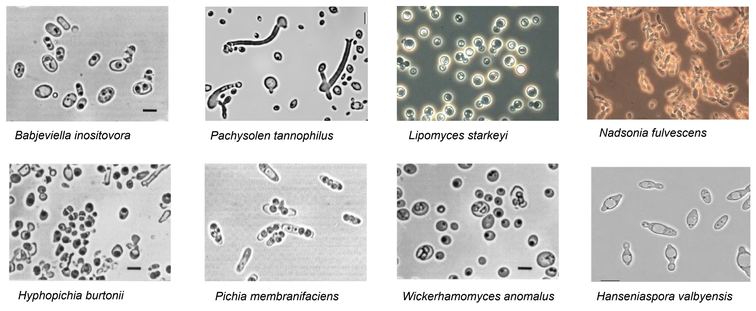
Significance

The largest fungal phylum, Ascomycota (ascomycetes), contains more than 60,000 described species and includes the budding yeasts (in Saccharomycotina) and fission yeasts (in Taphrinomycotina). Many of these yeasts have biotechnological, taxonomic and physiological interest. We present the genomes of 16 newly sequenced yeasts along with the genomes of several other previously published fungal genomes. We are mining these genomes to elucidate the biochemical, physiological, biotechnological, and bioconversion potential of an entirely new group of yeasts, which would expand our knowledge of the phylogenetic relationships of taxa in understudied lineages. Many of these understudied taxa are likely to have novel genes with biotechnological value.

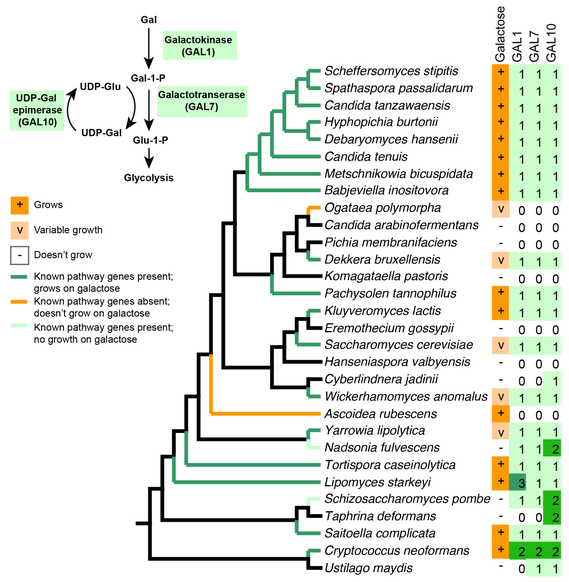
Conclusions

The genomes of 16 ascomycete yeasts, spanning two subphyla, are presented. Alternative CUG codon usage, based on analysis of CAG-tRNA structure, appears to be monophyletic. Galactose utilization is widespread and polyphyletic in yeasts spanning two phyla, with an inexact correlation between growth on galactose and the presence of known galactose metabolism genes, possibly reflecting inter-strain differences. Methylotrophy appears strictly dependent on a full complement of genes from the known methanol metabolism pathway.

We sequenced the genomes of 16 diverse ascomycete yeasts including:

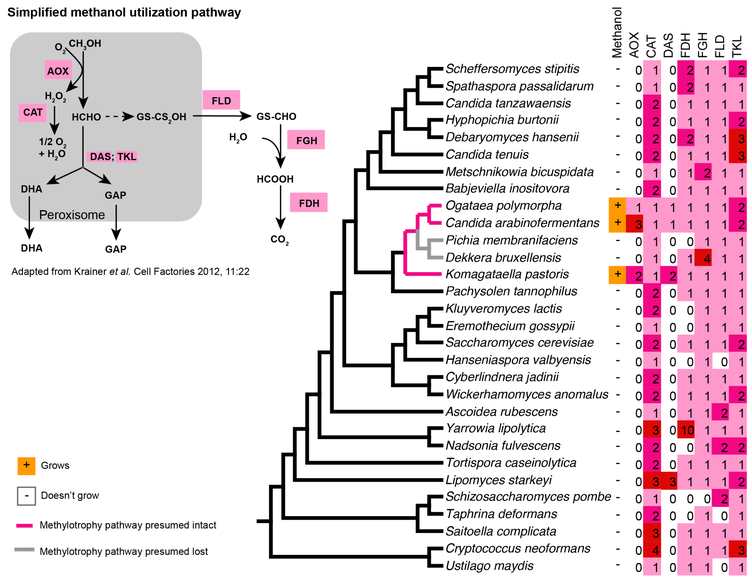


Galactose utilization. Galactose is a hexose sugar found in lignocellulose among other sources, which can be utilized by many yeasts, and in some cases, fermented into ethanol. The first three steps of galactose metabolism are catalyzed by GAL1 (galactokinase), GAL7 (galactose-1-phosphate uridyl transferase), GAL10 (UDP-glucose-4-epimerase); (Douglas and Hawthorne, Genetics 1966). In general, galactose utilization is widespread in the yeasts, including Taphrinomycotina and Basidiomycota, and is accompanied by known galactose utilization genes. However, some strains of *Ogataea polymorpha* and *Ascoidea rubescens* are reported to utilize galactose, while those we sequenced lack known galactose utilization genes. Moreover, *Nadsonia fulvescens* and *Schizosaccaromyces pombe* possess galactose utilization genes, yet appear not to grow on galactose. These anomalies may be the result of experimental error, misannotation, or differences among strains, but may also indicate our incomplete understanding of galactose utilization in the yeasts.



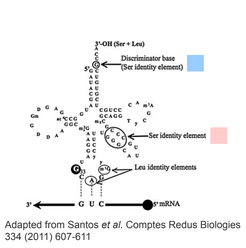
Enzymes were assigned using PRIAM (Claudel-Renard *et al*. NAR 2003). Growth profiles on methanol were taken from Kurtzman *et al*., The Yeasts: A Taxonomic Study, 5th ed.

Methylotrophy (methanol utilization). Several yeast species can metabolize methanol, including the newly-presented *Ogataea polymorpha* and *Candida arabinofermentans*. To investigate the evolution of methylotrophy, we mined the yeast genomes for the methanol pathway genes described in Krainer *et al*. Methyotrophy appears to have evolved once within the Saccharomycetales, and only the known methylotrophs contain a complete set of methylotrophy genes. Additionally, the distribution of methylotrophy genes on the phylogenetic tree implies that losses of the AOX, DAS, and FDH genes led to the loss of methylotrophy in *Pichia membranifaciens* and *Dekkera bruxellensis*.

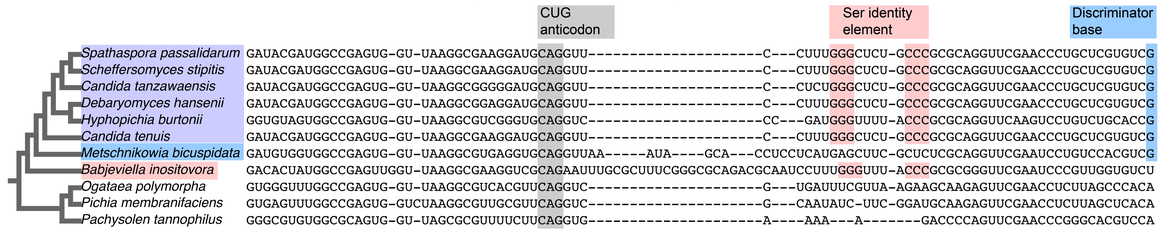


AOX: alcohol oxidase, CAT: catalase, DAS: dihydroxyacetone synthase, FDH: formate dehydrogenase, FGH: S-formylglutathione hydrolase, FLD: formaldehyde dehydrogenase, TKL: transketolase. Enzymes were assigned using PRIAM (Claudel-Renard *et al*. NAR 2003). Growth profiles on methanol were taken from Kurtzman *et al*., The Yeasts: A Taxonomic Study, 5th ed.

Alternative CUG codon usage in ascomycete yeasts. Although the genetic code is generally universal across bacteria and eukaryotes, some yeasts in the Saccharomycetales, collectively referred to as the 'CUG clade', translate CUG codons as Ser rather than Leu. The corresponding CAG-tRNAs have two conserved features: the Ser identity element, and the discriminator base. *Metschnikowia bicuspidata* and *Babjeviella inositovora*, basal to the rest of the CUG clade, each have only one of the features (discriminator base and Ser identity element, respectively). CAG-tRNAs from a closely related clade of yeasts in the Saccharomycetales lack either feature. The CAG-tRNA features we associate with CUG coding for Ser in Saccharomycetales yeasts appears to be monophyletic.

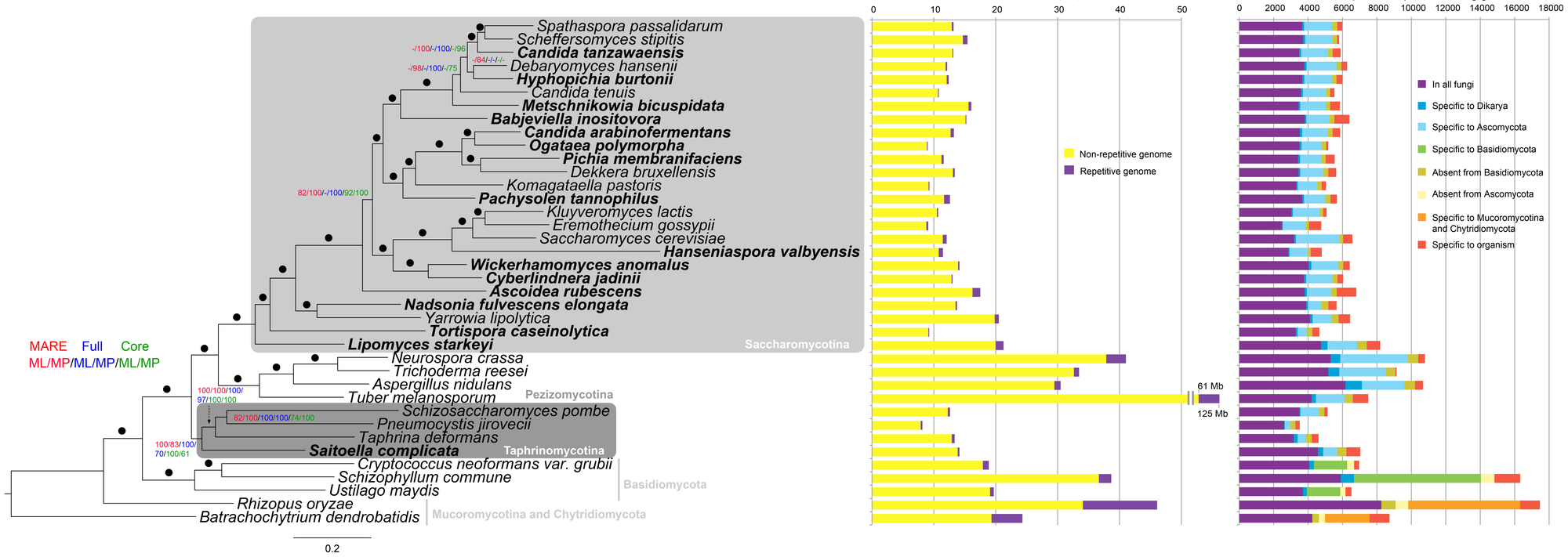


Adapted from Santos *et al.* Comptes Redus Biologies 334 (2011) 607-611



tRNAs were predicted from genomes with tRNAscan-SE (Lowe *et al.* 1997) and aligned with R-Coffee (Wilm *et al.* 1998).

Phylogeny and genome statistics of ascomycete yeasts with sequenced genomes. Phylogenetic tree inferred from the 364,126 aligned amino acid character containing MARE-filtered supermatrix under the maximum likelihood (ML) criterion and rooted with *Batrachochytrium dendrobatidis*. The branches are scaled in terms of the expected number of substitutions per site. Numbers on the branches are pairs of maximum-likelihood (left) and maximum-parsimony (right) bootstrap support values if larger than 60% from the MARE-filtered supermatrix (left), the full supermatrix (centre) and the core-genes supermatrix (right). Values larger than 95% are shown in bold; dots indicate branches with maximum support under all settings.



The 38 genome sequences were phylogenetically investigated using the DSMZ phylogenomics pipeline as previously described (Spring *et al*., 2010; Anderson *et al*., 2011; Göker *et al*., 2011; Abt *et al*., 2012, 2013; Breider *et al*., 2014; Frank *et al*., 2014; Scheuner *et al*., 2014) using NCBI BLAST (Altschul *et al*., 1997), OrthoMCL (Li *et al*., 2003), MUSCLE (Edgar, 2004), RASCAL (Thompson *et al*., 2003) and GBLOCKS (Talavera & Castresana, 2007) and MARE (Meusemann *et al*., 2010). That is, clusters of orthologs were generated using OrthoMCL, in-paralogs were removed, the remaining sequences were aligned with MUSCLE and filtered with RASCAL and GBLOCKS. Three distinct supermatrices were compiled, (i) all filtered alignments comprising at least four sequences; (ii) this "full" matrix cleaned from relatively uninformative genes and those with comparatively low coverage using MARE (Meusemann *et al*. 2010) under default values except that deleting organisms was disallowed; (iii) a core-genes matrix comprising only those genes present in all organisms. Maximum likelihood and maximum-parsimony trees were inferred from the concatenated alignments with RAxML (Stamatakis, 2006) and PAUP* (Swofford, 2002), respectively, as previously described (Spring *et al*., 2010; Anderson *et al*., 2011; Göker *et al*., 2011; Abt *et al*., 2012, 2013; Breider *et al*., 2014; Frank *et al*., 2014; Scheuner *et al*., 2014).

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.